

Penalized EM algorithm and copula skeptic graphical models for inferring networks for mixed variables

Fentaw Abegaz and Ernst Wit

Johann Bernoulli Institute of Mathematics and Computer Science
University of Groningen

Abstract

In this article, we consider the problem of reconstructing networks for continuous, binary, count and discrete ordinal variables by estimating sparse precision matrix in Gaussian copula graphical models. We propose two approaches: ℓ_1 penalized extended rank likelihood with Monte Carlo Expectation-Maximization algorithm (copula EM glasso) and copula skeptic with pair-wise copula estimation for copula Gaussian graphical models. The proposed approaches help to infer networks arising from nonnormal and mixed variables. We demonstrate the performance of our methods through simulation studies and analysis of breast cancer genomic and clinical data and maize genetics data.

Keywords: Gaussian copula; ℓ_1 penalized maximum likelihood; Gaussian graphical models; EM algorithm; Extended rank likelihood; Nonparanormal skeptic; Copula skeptic.

1 Introduction

The aim of this article is to formulate an inference approach for the analysis of high dimensional data that involves mixed variables of continuous, binary and ordered categorical types using graphical models. In particular, we focus on model estimation and identification of undirected graph structure for Gaussian graphical models for high dimensional datasets. We base our inference procedure on the EM algorithm and pair-wise copula estimation with ℓ_1 penalized extended rank likelihood.

Toward the study of mixed variables determination of their joint distribution is the main challenge. The seminal work of Sklar (1959) that formally introduced the notion of copula provide the theoretical framework, in which a joint probability distribution can be represented by its univariate marginal distributions and a copula function. As a result multivariate association, which is fully described by the copula function, can be modeled separately from the univariate marginal distributions.

In copula modeling, Genest, Ghoudi, and Rivest (1995) developed a popular semi-parametric estimation or “rank likelihood” based estimation, in which the association among the variables are represented by a parametric copula model but the marginals are

treated as nuisance parameters and estimated nonparametrically. The resulting semi-parametric estimators are well-behaved for continuous data but fail for discrete data, for which the distribution of the ranks depends on the univariate marginal distributions, making them somewhat inappropriate for the analysis of mixed continuous and discrete data (Hoff, 2007). To remedy this, Hoff (2007) propose the extended rank likelihood, which is a type of marginal likelihood that does not depend on the marginal distributions of the observed variables. Under the extended rank likelihood approach, the ranks are free of the nuisance parameters (or marginal distributions) of the discrete data. This makes the extended rank likelihood approach more suited for the determination of graphical models in the mixed variable setting and avoids the difficult problem of modeling marginal distributions (Dobra and Lenkoski, 2011).

The extended rank likelihood estimation is implemented for the study of association among mixed variables under a Bayesian framework by Hoff (2007) and further studied in the graphical model setting by Dobra and Lenkoski (2011) using Bayesian model averaging approach for graph identification and estimation in copula Gaussian graphical models. Since the marginals are treated as nuisance parameters, the parameter of interest for estimation is the correlation matrix or the precision matrix, i.e. the inverse of the correlation matrix in case of a Gaussian copula. Ambroise, Chiquet, and Matias (2009) raised their concern on the challenging task involved in the Bayesian framework to construct priors on the set of precision or concentration matrices. In this article we propose an alternative approach that consider the extended rank likelihood under l_1 penalized maximum likelihood setting with the Expectation-Maximization (EM) algorithm for high-dimensional inference based on graphical models. This approach is referred to as copula EM glasso.

On the other hand, Liu et al. (2012) considered graphical modeling for binary and continuous variables using nonparanormal distributions and glasso algorithm of Friedman, Hastie, and Tibshirani (2008). In particular, in their nonparanormal skeptic approach, they considered the rescaled empirical distribution transformation of data (with or without truncation and monotone transformation) to compute correlation matrix based on nonparametrically estimated pairwise rank correlations. We observe that the use of one step glasso algorithm makes their approach computationally efficient for high dimensional setting. Further we note that rank correlations such as Kendall's tau and Spearman's rho are directly related to bivariate copula models. Through these relationships and upon carefully selected bivariate copulas more accurate estimation of the rank correlations can be achieved. Thus, we extend the paranonnormal skeptic approach for rank correlations computed from bivariate parametric copulas. This approach is referred to as copula skeptic glasso.

We apply the proposed approaches to breast cancer genomic and clinical data. Breast cancer is the leading cause of death among women in the world and represents a significant health problem. Multiple factors like age, diet, obesity, parity, age at first childbirth, oral contraceptives, exogenous estrogens, genetics, environment, geographic location influence the development of breast cancer. However, the majority of the cases in breast cancer is always due to genetic abnormalities. At present, only small numbers

of accurate prognostic and predictive factors are used clinically for managing the patients with breast cancer (Kumar et al., 2012). In the last few decades knowledge of breast cancer grade determined by Nottingham prognostic index (NPI) has been very helpful to decide on the most effective treatments. Moreover, microarray-based gene expression profiling has been used extensively to characterize the transcriptome of breast cancer, resulting in the identification of new molecular subtypes and markers or signatures of potential therapeutic and prognostic importance (Ringnér et al., 2011). Inclusion of such treatment predictive markers considerably improved breast cancer treatment decisions. To further tailor treatment for individual patients, identification of additional clinical and genetic markers is required.

Genomic DNA copy number alterations, i.e., amplifications or deletions, are key genetic events in the development and progression of breast cancers. Gene copy number changes can be determined on a gene-by-gene basis using microarrays. A genome-wide microarray comparative genomic hybridization (CGH) is used to analyse the pattern of DNA copy number alteration with the aim to study the relationship between DNA amplification and deletion patterns and severity of breast cancer as measured by several clinical indicators on patients. Details of the experiment is discussed in Jensen et al. (2009). The data from the breast cancer experiment include 296 variables of which 287 are genes and 9 are clinical variables obtained from 106 breast cancer patients. The genomic and clinical variables are mixed measurements of continuous, binary and ordered categorical types, see details in Section 4.

We also considered a second application of our approach on a very high dimensional setting on data from maize genetic nested association mapping population that has been analysed and discussed by McMullen et al. (2009). The data set includes 1106 SNP loci or genetic markers and 4699 replicates. Our objective is to obtain a sparse representation of potential trans-acting genetic markers that may provide information for a better understanding of the molecular basis of phenotypic variation. This helps to improve agricultural efficiency and sustainability.

The rest of this article is organized as follows. Section 2.1 provides a brief description of Gaussian copula modeling aspects related to continuous, binary, count and ordinal variables. Section 3.1 formulates the copula based EM algorithm with l_1 penalized likelihood estimation and the copula skeptic glasso with pair-wise copula selection. It also discusses the selection of tuning parameter in case of EM formulation of glasso. Section 4 demonstrates performance of the proposed approach using simulation studies and the analysis of high dimensional data on breast cancer and maize genetic properties. We close with a concluding remark in Section 5.

2 Copula Graphical models

2.1 Gaussian copula graphical models

Graphical models are efficient tools for studying multivariate distributions through a compact, graphical representation of the joint probability distribution of the underlying

random variables. The treatment of graphical models simplifies significantly, when one focuses on normally distributed variables. Let the random vector $Y = (Y_1, \dots, Y_p)^T$ be assumed to be Gaussian with a positive-definite covariance matrix Σ of dimension $p \times p$. A graphical model $G = (V, E)$, where V corresponds to the set of nodes with p elements and $E \subset V \times V$ of ordered pairs of distinct nodes called the edges of G , for $\mathcal{N}_p(0, \Sigma)$ is called a Gaussian graphical model, if on the graph G , the edges E represent conditional dependence among the random variables. Absence of an edge between any pair of random variables or zero value of a precision matrix, $\Theta = \Sigma^{-1}$, corresponds to conditional independence of the two random variables given the remaining ones.

In practice, we encounter both discrete and continuous variables that may not be Gaussian. Thus, the assumption of multivariate normal distribution would be too restrictive. To relax the normality requirement, we use the copula framework to construct multivariate distributions for arbitrary marginals as discussed in Section 1 above. In order to use the properties of the Gaussian graphical model, we consider the Gaussian copula. The Gaussian copula with correlation matrix Γ of dimension $p \times p$ having $p(p-1)/2$ free parameters is given by:

$$C(u_1, \dots, u_p \mid \Gamma) = \Phi_p(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_p) \mid \Gamma),$$

and the corresponding Gaussian copula-based distribution function is

$$H(y_1, \dots, y_p \mid \Gamma, F_1, \dots, F_p) = \Phi_p(\Phi^{-1}(F_1(y_1)), \dots, \Phi^{-1}(F_p(y_p)) \mid \Gamma). \quad (2.1)$$

Here $\Phi(\cdot)$ represents the CDF of the standard normal distribution and $\Phi_p(\cdot \mid \Gamma)$ is the CDF of multivariate normal distribution $\mathcal{N}_p(0, \Gamma)$.

We note that under the Gaussian copula the correlation matrix Γ is the matrix of correlation coefficients among the transformed variables $Z_j = \Phi^{-1}(F_j(y_j))$, $j = 1, \dots, p$, which represent the maximum pairwise correlations among the Y_j s, $j = 1, \dots, p$. If the univariate marginal distributions are normal, then entries of the correlation matrix represent exactly the pairwise correlation coefficients of the variables.

Proposition 2.1 *Let Γ be a positive-definite matrix, such that $Z \sim \mathcal{N}_p(0, \Gamma)$ is a graphical model with respect to graph $G = (V, E)$. Then the continuous variable Y , defined via 2.1, is also a graphical model with respect to G . In particular, the precision matrix $\Theta = \Gamma^{-1}$ represents the conditional independence among the observed variables Y_j s.*

Proof 2.1 *This is as a result of the invariance property of conditional independence relation over equivalent probability measures as shown in Van Putten and Van Schuppen (1985, Theorem 3.6).*

We now focus on graphical modeling for observed variables Y of mixed type, i.e. in the case that they represent a collection of continuous, binary, ordinal or count variables. Suppose the j -th variable Y_j has marginal distribution F_j with its pseudo-inverse F_j^{-1} . In copula modeling, the marginals are treated as nuisance parameters and estimated nonparametrically mainly using the rescaled empirical distribution: $\hat{F}_j(y) =$

$\frac{1}{n+1} \sum_{i=1}^n \mathcal{I}\{Y_{ji} \leq y\}$, $j = 1, \dots, p$. A copula graphical model, discussed above, can be constructed by introducing a vector of latent variables $Z \sim \mathcal{N}_{\sqrt{}}(0, \Gamma)$ that are related to the observed variables Y as $Y_j = \hat{F}_j^{-1}(\Phi(Z_j))$, $j = 1, \dots, p$. In the case of mixed variables, the graphical structure, i.e. the conditional independence implied by the graph structure, is assumed to hold exclusively on the latent variable Z .

The aim of the inference procedure is to infer the graphical structure G , defined by the latent variable Z . Though the Z s are not observable, the observed Y_j s do provide a limited amount of information about them. Since the \hat{F}_j s are nondecreasing, observing $Y_{k_1} < Y_{k_2}$ implies that $Z_{k_1} < Z_{k_2}$. More generally, observing the n -dimensional vector Y_i tells us that Z_i lies in

$$D(Y_i) = \{z \in \mathbb{R}^n | a_i(Y_{ij}) < Z_j \leq b_i(Y_{ij})\}, \quad (2.2)$$

where $a_i(y) = \inf\{z | \hat{F}_j^{-1}(\Phi(z)) = y\}$ and $b_i(y) = \sup\{z | \hat{F}_j^{-1}(\Phi(z)) = y\}$. In fact, for every ordinal Y_j , we can identify a set of thresholds $\tau_j = (\tau_{j0}, \tau_{j1}, \dots, \tau_{jn_j})$ with

$$-\infty = \tau_{j0} < \tau_{j1} < \dots < \tau_{jn_j} = \infty,$$

such that for some ordered set of values $\{c_{j1} < \dots < c_{jn_j}\}$,

$$Y_j = \sum_{r=1}^{n_j} c_{jr} \times I\{\tau_{j,r-1} < z_j \leq \tau_{jr}\}.$$

It follows that the mapping of the ordered values of Y_j into some defined intervals $(\tau_{jr}, \tau_{j,r+1}]$ of the latent variable Z_j relies on the following relationship

$$\tau_{jr} = \Phi^{-1}(\hat{F}_j(c_{jr})), \quad r = 1, \dots, n_j - 1.$$

The collection of these intervals is the set $D(Y) = D(Y_1, \dots, Y_n)$ in (2.2). In case of missing observations, we consider data missing in this study as missing completely at random so that the missing values are easily determined from the latent variable distribution defined on the interval $(-\infty, \infty)$. Then the occurrence of event $Z \in D(Y)$ is taken as the observed data to infer about the copula parameter or the graph structure separately from the marginal distributions. Such inference approach is similar to the extended rank likelihood in Hoff (2007) and copula Gaussian graphical modeling in Dobra and Lenkoski (2011).

3 Sparse inference methods

3.1 Copula EM GLASSO approach

The Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) is a popular method for maximum likelihood estimation in the case of incomplete data, which naturally occur in our setting as a result of the latent nature of Z as discussed

in the previous section. In this section we consider EM algorithm with penalized likelihood. Green (1990) studied convergence properties of the EM algorithm for penalized likelihood.

The marginal likelihood of Y where we consider F_1, \dots, F_p as nuisance parameters; see also Hoff (2007), is

$$L_Y(\Theta) = \int_{D(Y)} p(z \mid \Theta) dz \quad (3.1)$$

For large sample sizes the precision matrix Θ can be estimated by maximizing the log-likelihood $l(\Theta)$ as a function of Θ . Whereas for high-dimensional data ($n \ll p$), we add an l_1 -norm penalty to encourage sparsity in the precision matrix and the identification of the underlying graph. The l_1 penalized log-likelihood is given by

$$\ell_\lambda(\Theta) = \log L_Y(\Theta) - \lambda \|\Theta\|_1, \quad (3.2)$$

where the scalar parameter $\lambda \geq 0$ controls the size of the penalty.

Due to the complexity of maximizing the marginal log-likelihood $l_Y(\Theta)$ in (3.1) and (3.2), we employ the EM algorithm. We have discussed in the previous section that the observed data Y provide some information on the latent variables Z such that the occurrence of the event $Z \in D(Y)$ is taken as the observed data to infer on Θ . We now recall from standard EM algorithm setting that the loglikelihood of the observed data can be expressed as

$$\ell(\Theta) = Q(\Theta \mid \Theta^{(m)}) - H(\Theta \mid \Theta^{(m)}), \quad (3.3)$$

where $\Theta^{(m)}$ an estimate of Θ from the previous step of the algorithm,

$$Q(\Theta \mid \Theta^{(m)}) = E \left[\log L_{Z, Z \in D(Y)}(\Theta) \mid Z \in D(Y), \Theta^{(m)} \right], \quad (3.4)$$

and

$$H(\Theta \mid \Theta^{(m)}) = E \left[\log L_{Z \mid Z \in D(Y)}(\Theta) \mid Z \in D(Y), \Theta^{(m)} \right]. \quad (3.5)$$

The penalized log-likelihood takes the form

$$\ell_\lambda(\Theta) = Q(\Theta \mid \Theta^{(m)}) - H(\Theta \mid \Theta^{(m)}) - \lambda \|\Theta\|_1, \quad (3.6)$$

such that

$$\Theta_\lambda^{(m+1)} = \operatorname{argmax}_\Theta \left\{ Q(\Theta \mid \Theta^{(m)}) - H(\Theta \mid \Theta^{(m)}) - \lambda \|\Theta\|_1 \right\}. \quad (3.7)$$

Further, from standard EM approach, $H(\Theta \mid \Theta^{(m)}) \leq H(\Theta^{(m)} \mid \Theta^{(m)})$ for any Θ in the parameter space. Thus, obtaining an updated estimate of the parameter by maximizing the l_1 penalized log-likelihood in (3.7) reduces to

$$\Theta_\lambda^{(m+1)} = \operatorname{argmax}_\Theta \left\{ Q(\Theta \mid \Theta^{(m)}) - \lambda \|\Theta\|_1 \right\}. \quad (3.8)$$

The EM optimization strategy alternates iteratively between the E-step, computing conditional expectation of the complete log-likelihood $Q(\Theta \mid \Theta^{(m)})$ and the M-step, maximizing $Q(\Theta \mid \Theta^{(m)})$, with a sparsity penalty $\lambda \|\Theta\|_1$, over Θ .

E-step: The complete data likelihood depends on the joint distribution of $(Z, Z \in D(Y))$ given by

$$p(Z, Z \in D(Y) \mid \Theta) = \begin{cases} p(Z \mid \Theta) & Z \in D(Y) \\ 0 & Z \notin D(Y) \end{cases}$$

where $p(Z \mid \Theta)$ is the multivariate normal density with mean zero and variance $\Sigma = \Theta^{-1}$. Then the complete log likelihood of $(Z, Z \in D(y))$ for a random sample of size n after ignoring constants with respect to Θ is given by

$$\begin{aligned} l_Z(\Theta) &= \sum_{i=1}^n \log(p(Z_i \mid \Theta)) I_{\{Z_i \in D(Y_i)\}} \\ &= -\frac{np}{2} \log(2\pi) + \frac{n}{2} \log \det(\Theta) - \frac{1}{2} \sum_{i=1}^n Z_i^T \Theta Z_i I_{\{Z_i \in D(Y_i)\}} \end{aligned} \quad (3.9)$$

Using the complete log likelihood in (3.9), it follows that

$$\begin{aligned} Q(\Theta \mid \Theta^{(m)}) &= E \left[l_Z(\Theta) \mid Z \in D(Y), \Theta^{(m)} \right] \\ &= \frac{n}{2} \left\{ \log \det(\Theta) - \frac{1}{n} \sum_{i=1}^n \left(E \left[Z_i^T \Theta Z_i \mid Z_i \in D(Y_i), \Theta^{(m)} \right] \right) \right\} \\ &= \frac{n}{2} \left\{ \log \det(\Theta) - \text{Tr} \left(\Theta \frac{1}{n} \sum_{i=1}^n E \left[Z_i Z_i^T \mid Z_i \in D(Y_i), \Theta^{(m)} \right] \right) \right\} \\ &= \frac{n}{2} \{ \log \det(\Theta) - \text{Tr}(\Theta \bar{R}) \}, \end{aligned} \quad (3.10)$$

where Tr stands for the trace of a matrix and \bar{R} is the expected empirical covariance function of the latent variables given $Z \in D(Y)$:

$$\bar{R} = \frac{1}{n} \sum_{i=1}^n E \left[Z_i Z_i^T \mid z_i \in D(Y_i), \Theta^{(m)} \right],$$

where

$$\begin{aligned} E \left[Z_i Z_i^T \mid Z_i \in D(Y_i), \Theta^{(m)} \right] &= E \left[Z_i \mid Z_i \in D(y_i), \Theta^{(m)} \right] E \left[Z_i \mid Z_i \in D(y_i), \Theta^{(m)} \right]^T \\ &\quad + \text{cov} \left[Z_i \mid Z_i \in D(y_i), \Theta^{(m)} \right] \end{aligned} \quad (3.11)$$

Note that the conditional latent random variable $\{Z \mid Z \in D(Y)\}$ follows a truncated multivariate normal distribution. Analytical expressions for the computations of moments for truncated multivariate normal distribution are given in Wilhelm and Manjunath (2010) and references therein. However, due to the computational complexity,

obtaining analytical solutions is only feasible for very few variables. Another approach is to simulate a large sample from the truncated multivariate normal distribution and calculate the sample conditional covariance matrix and sample conditional mean to estimate $E[Z_i Z_i^T | Z_i \in D(y_i), \Theta^{(m)}]$ using (3.11).

Alternatively, towards a computational efficient approach instead of mapping all mixed variables to a latent space as discussed above we may partition the mixed variables into two as continuous denoted by Y_c , and ordered variables that includes ordinal, binary and counts denoted by Y_o . We also partition the correlation matrix along with the variables grouping as

$$\Sigma = \begin{bmatrix} \Sigma_{cc} & \Sigma_{co} \\ \Sigma_{oc} & \Sigma_{oo} \end{bmatrix}.$$

We then take $Z = (Z_c, Z_o) \sim N(0, \Theta)$, where $\Theta = \Sigma^{-1}$ with $Z_c = \Phi^{-1}(\hat{F}(Y_c))$ is transformed normal scores of observed continuous variables using the rescaled empirical distribution based on the natural Gaussian copula semiparametric approach and $Z_o \in D(Y_o)$ is the latent normal score corresponding to the ordered observed variables Y_o obtained in a similar way as discussed in Section 2.1.

With a similar argument as above the complete data likelihood depends on the joint distribution: $p(Z_c, Z_o, Z_o \in D(Y_o) | \Theta) = p(Z_c, Z_o)$, for $Z_o \in D(Y_o)$. Such that the complete data loglikelihood after ignoring constants is given by

$$\begin{aligned} l_{Z_c, Z_o}(\Theta) &= \sum_{i=1}^n \log(p(Z_{c_i}, Z_{o_i} | \Theta)) I_{\{Z_{o_i} \in D(Y_{o_i})\}} \\ &= \frac{n}{2} \log \det(\Theta) - \frac{1}{2} \sum_{i=1}^n [Z_{c_i}, Z_{o_i}]^T \Theta [Z_{c_i}, Z_{o_i}] I_{\{Z_{o_i} \in D(Y_{o_i})\}} \end{aligned} \quad (3.12)$$

Using this complete data log-likelihood and after ignoring constants with respect to Θ , it follows that

$$\begin{aligned} Q(\Theta | \Theta^{(m)}) &= E_{Z_o} [l_{Z_c, Z_o}(\Theta) | Y_c, Z_o \in D(Y_o), \Theta^{(m)}] \\ &= \frac{n}{2} \{\log \det(\Theta) \end{aligned} \quad (3.13)$$

$$\begin{aligned} &\quad - \text{Tr} \left(\Theta \frac{1}{n} \sum_{i=1}^n E_{Z_o} [[Z_{c_i}, Z_{o_i}] [Z_{c_i}, Z_{o_i}]^T | Y_{c_i}, Z_{o_i} \in D(Y_{o_i}), \Theta^{(m)}] \right) \} \\ &= \frac{n}{2} \left\{ \log \det(\Theta) - \text{Tr} (\Theta \tilde{R}) \right\}, \end{aligned} \quad (3.14)$$

where

$$\tilde{R} = \frac{1}{n} \sum_{i=1}^n E_{Z_o} [[Z_{c_i}, Z_{o_i}] [Z_{c_i}, Z_{o_i}]^T | Y_{c_i}, Z_{o_i} \in D(Y_{o_i}), \Theta^{(m)}]. \quad (3.15)$$

An estimate of \tilde{R} can be obtained after evaluating the expectations as follows.

$$\begin{aligned} E_{Z_o} [Z_{c_i} Z_{c_i}^T | Y_{c_i}, Z_{o_i} \in D(Y_{o_i}), \Theta^{(m)}] &= Z_{c_i} Z_{c_i}^T \\ E_{Z_o} [Z_{c_i} Z_{o_i}^T | Y_{c_i}, Z_{o_i} \in D(Y_{o_i}), \Theta^{(m)}] &= Z_{c_i} \hat{Z}_{o_i}^T \\ E_{Z_o} [Z_{o_i} Z_{o_i}^T | Y_{c_i}, Z_{o_i} \in D(Y_{o_i}), \Theta^{(m)}] &= \hat{Z}_{o_i} \hat{Z}_{o_i}^T + \Theta_{oo}^{(m)-1}. \end{aligned}$$

where $\hat{Z}_{o_i}^T = E_{Z_o} [Z_{o_i} | Z_{o_i} \in D(Y_{o_i}), \Theta^{(m)}] - \Theta_{oo}^{(m)-1} \Theta_{oc}^{(m)} Z_{c_i}$, is a conditional expectation defined on the distribution $p(Z_o | Y_c)$ for $Z_o \in D(Y_o)$.

M-step: This involves updating the parameter Θ using the l_1 penalized log-likelihood given by

$$\Theta_\lambda^{(m+1)} = \operatorname{argmax}_\Theta \left\{ Q(\Theta | \Theta^{(m)}) - \lambda \|\Theta\|_1 \right\}. \quad (3.16)$$

We next substitute the Q function by (3.10) or (3.14) from the E-step to obtain

$$\Theta_\lambda^{(m+1)} = \operatorname{argmax}_\Theta \left\{ \log \det(\Theta) - \operatorname{Tr}(\Theta \bar{R}) - \lambda \|\Theta\|_1 \right\}. \quad (3.17)$$

The maximization problem in (3.17) takes the form of l_1 penalized likelihood for Gaussian graphical models and computation is done efficiently using the graphic lasso algorithm (Friedman et al., 2008). This algorithm is fast and allows the re-use of the estimate under one value of the tuning parameter as a “warm” start for the next value. The determination of a value for λ in case of penalized inference with EM algorithm is discussed in Section 3.3.

Remark 3.1 *Setting the penalty parameter $\lambda = 0$ results in the unpenalized maximum likelihood estimate which can be considered as an alternative to the Bayesian approach discussed in (Hoff, 2007).*

3.2 Copula skeptic GLASSO

The copula EM glasso approach discussed in the previous section, though it is a natural approach, it is computationally expensive, since it calls MCMC in the E-step and glasso in the M-step. In particular, in a very high dimensional setting the computational issue requires further attention. One approach is to seek a one-to-one mapping of the observed and latent variables that avoids the Monte Carlo EM algorithm resulting from a one-to-many mapping. The semiparametric copula approach of estimating marginals through the rescaled empirical distribution is a one-to-one mapping or transformation of the observed data. Instead of directly using the transformed data to estimate Θ , a sample correlation matrix can be compute from pairwise rank correlations. In this regard, Liu et al. (2012) considered a nonparanormal skeptic approach to obtain sparse estimates of Θ for binary and continuous variables using one step glasso based on the estimated correlation matrix.

We note that the use a one step glasso approach is computationally efficient. Further, we note that rank correlations like Kendall’s tau and Spearman’s rho can be better

approximated by a carefully chosen parametric bivariate copula model that takes in to account the underlying bivariate dependence structure. The vast literature on copulas deals with bivariate copula models and has demonstrated their potential to capture various types of dependence structure.

It is known, for example, that the population version of Kendall's tau is related to parametric copula models parametrized by γ_{ij} via

$$\tau_{ij} = 4 \int_0^1 \int_0^1 C(u, v \mid \theta_{ij}) dC(u, v \mid \gamma_{ij}) - 1.$$

For commonly used copula models, there is closed form representation of the Kendall's tau using the bivariate copula parameter, see for example Nelsen (2006). An estimate of Kendall's tau is obtained using

$$\hat{\tau}_{ij} = \begin{cases} 4 \int_0^1 \int_0^1 C(u, v \mid \hat{\theta}_{ij}) dC(u, v \mid \hat{\gamma}_{ij}) - 1 & \text{for } i \neq j \\ 1 & \text{for } i = j \end{cases}$$

or its closed form representation. Further Kendall's tau is related to the correlation coefficient, Γ , by

$$\hat{\Gamma}_{ij} = \begin{cases} \sin\left(\frac{\pi}{2} \hat{\tau}_{ij}\right) & \text{for } i \neq j \\ 1 & \text{for } i = j \end{cases}$$

Then to obtain sparse estimates, glasso can be implemented that uses the estimated correlation matrix $\hat{\Gamma}$ in the direct optimization of the objective function:

$$\hat{\Theta}_\lambda = \operatorname{argmax}_{\Theta} \left\{ \log \det(\Theta) - \operatorname{Tr}(\Theta \hat{\Gamma}) - \lambda \|\Theta\|_1 \right\}.$$

3.3 Model selection

For high dimensional data, the empirical covariance matrix is singular and poses computational problems. However, our l_1 penalized approach guarantees with probability one a positive definite precision matrix with the additional property of being sparse. Note that sparseness refers to the property that all parameters that are zero are actually estimated as zero with probability tending to one. This helps to assess conditional independence based on entries of the precision matrix (Dempster, 1972).

Under the l_1 penalized maximum likelihood setting the sparsity of the estimated precision matrix is controlled by the penalty parameter λ in (3.17). We follow information based criteria in order to obtain reasonably sparse precision matrix. One could also use cross-validation for the choice of λ which we have not consider in this article.

We now consider (3.3) that suggests the log likelihood of the observed data can be computed at EM convergence, see for example Ibrahim et al. (2008). Let the estimate $\hat{\Theta}_\lambda$ is obtained at EM convergence for a given value of λ . The log likelihood of the observed data is

$$\log L_Y(\hat{\Theta}_\lambda) = Q(\hat{\Theta}_\lambda \mid \hat{\Theta}_\lambda) - H(\hat{\Theta}_\lambda \mid \hat{\Theta}_\lambda).$$

Thus a model selection criterion is defined by

$$\begin{aligned} IC(\lambda) &= -2\log L_Y(\hat{\Theta}_\lambda) + pen(\hat{\Theta}_\lambda) \\ &= -2Q(\hat{\Theta}_\lambda | \hat{\Theta}_\lambda) + 2H(\hat{\Theta}_\lambda | \hat{\Theta}_\lambda) + pen(\hat{\Theta}_\lambda), \end{aligned}$$

where $pen(\hat{\Theta}_\lambda)$ refers to a penalty term. Different forms of $pen(\hat{\Theta}_\lambda)$ lead to different model selection criteria. Let d denotes the number of non-zero upper or lower off-diagonal elements of $\hat{\Theta}_\lambda$. Thus we define AIC and BIC as follows:

$$\begin{aligned} AIC(\lambda) &= -2Q(\hat{\Theta}_\lambda | \hat{\Theta}_\lambda) + 2H(\hat{\Theta}_\lambda | \hat{\Theta}_\lambda) + 2d, \\ BIC(\lambda) &= -2Q(\hat{\Theta}_\lambda | \hat{\Theta}_\lambda) + 2H(\hat{\Theta}_\lambda | \hat{\Theta}_\lambda) + \log(n)d. \end{aligned}$$

Then we choose the optimal value of the penalty parameter as the one that minimizes these criteria on a grid of candidate values for λ .

4 Analysis of data

4.1 Simulations

We carried out simulation studies with a variety of data structures to compare how well competing methods recover the true graph structure. Though our EM approach is computationally expensive, we noticed that in our simulations the EM algorithm converges very quickly with a maximum of ten iterations and 100 MCMC samples for hundreds of variables.

For the purpose of comparison we considered the following approaches:

1. Proposed copula EM glasso (CopulaEM).
2. Proposed copula skeptic glasso (CopulaTau)
3. Nonparanormal normal-score based estimation with truncation presented in Liu et al. (2012) (NPNscore)
4. Nonparanormal skeptic using Kendall's tau presented in Liu et al. (2012) (NPNtau)

In our simulation we consider sample sizes ($n=200$) and number of variables($p=100$) which are of mixed types that include binary(10), ordinal(10), count (10), nonnormal (eg. Chisquare (10)), and the remaining 60 are normal variables with outliers (none , 1% , 20%). In case of outliers, observations are replaced by a value either 5 or -5 with probability 0.6, see also Liu et al. (2012). ROC curves are used to compare performance of the different approaches in recovering the true graph.

Figures 1 and 2 displays ROC curves based on averages of true positive rates and false positive rates computed from 100 times repeated simulations at each of 10 grid points of the tuning parameter. For mixed data with no and low level outliers, we see that the difference in the performance of recovering the true graph based on the various

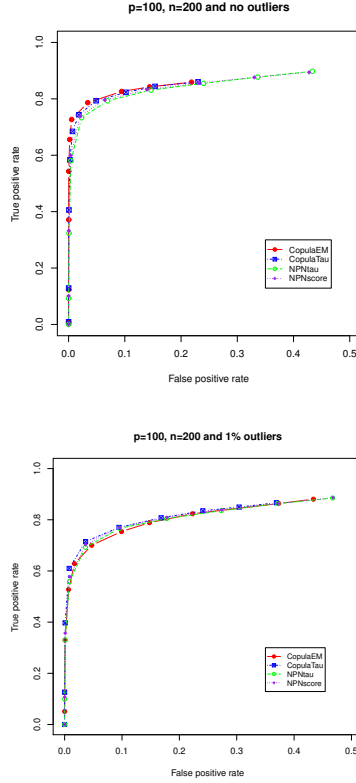


Figure 1: *ROC curves comparing various methods in recovering true graph structure for $n = 200$ and $p = 100$ in case of no and low level of outliers: our proposed approaches “CopulaEM” (copula EM glasso) and “CopulaTau” (copula skeptic) perform comparably to that of “NPNscore” (normal-score nonparanormal estimator) and “NPNtau” (nonparanormal skeptic)*

methods is negligible though our copula EM glasso shows slightly better performance in case of no outliers.

In case of high level of outliers with mixed variables, the performance of the proposed copula skeptic and nonparanormal skeptic are comparable but the proposed copula skeptic outperforms the nonparanormal skeptic. This suggests that a careful choice of parametric bivariate copulas results in better performance over the nonparametric approaches.

4.2 Applications

4.2.1 Breast cancer data

In this section, we return to the motivation of our methodological development and apply the proposed Copula EM glasso approach to the breast cancer data, which we introduced in Section 1. The breast cancer experiment is a clinical study of DNA amplification and deletion patterns, using microarray technology. Its aim is to study the relationship

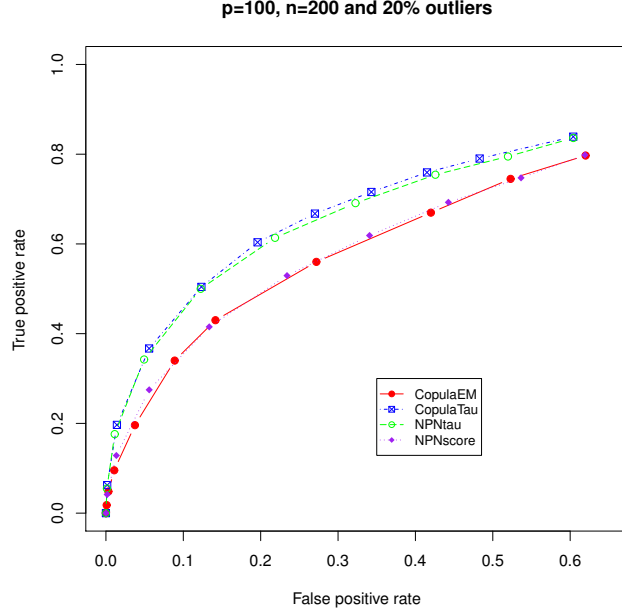


Figure 2: ROC curves comparing various methods in recovering true graph structure for $n = 200$ and $p = 100$ in the presence of high level of outliers: “CopulaTau” performs better than “NPNtau” while both out perform “CopulaEM” and “NPNscore”

between DNA amplification and deletion patterns (rather than gene expression) and the severity of the breast cancer, as measured by several clinical indicators on the patients. The data from the breast cancer experiment include 287 selected genes and 9 clinical variables obtained from 106 breast cancer patients. A brief description of clinical and genomic variables included in this study are presented in Table 1.

In the breast cancer study, missing data rates among each of the gene amplification variables was less than 3%. This could be that in microarray experiments it happens frequently that some part of the array could be damaged and resulted in some data to be excluded from consideration. Similarly, the missing data rates for the clinical variables were between 5% and 20%, respectively.

In this study, we express the relationship between breast cancer survival, genomic and clinical variables as a series of conditional dependencies. We applied the proposed Copula EM glasso described in this paper that internally samples missing observations. The BIC criterion resulted in an optimal penalty value of $\lambda = 0.15$. A subnetwork of the complex dependence pattern among the observed variables induced by the underlying multivariate normal latent variables is displayed in Figure 3. This subnetwork includes only links among the clinical and genomic variables.

As can be seen from Figure 3, breast cancer death is related to clinical variables (NPI score, Grade and size of breast cancer tumors) and markers like SHGC4-207 and 10QTEL24. As expected the NPI is directly related to breast cancer tumor size, cancer

Table 1: List of genomic and clinical variables obtained from the breast cancer experiment. The aim is to find the underlying conditional dependence structure between these binary, count, ordered categorical, and continuous clinical and genomic variables.

age.at.diagnosis	age at diagnostics (in years)	continuous
size	size of breast tumour (in mm)	continuous
survival status	died due to breast cancer	binary
grade	grade of breast cancer: 1 (low) to 3 (high)	ordered categorical
nodal stage	lymp nodes involved	count
NPI	NPI score	continuous
ER	ER status: positive or negative	binary
hist	Histology: Ductal or others	binary
Ther	Therapy: Hormone or others	binary
genes	gene amplification/deletion	continuous

grade and nodes involved. The higher the values on the clinical variables the more aggressive the breast cancer and higher chance of death due to breast cancer.

Further we see that the NPI score and the three clinical variables are related to the amplification or deletion of genes, for example, BRCA1, RPS6KB1, ABL1, BMI1, CREBBP, STK6 (STK15), VHL, CTSB, PDGRL, GARP, ATM and PIM1. These findings are consistent with the literature that revealed these genes are associated with the risk and progression of breast cancer, see for example, Bärnlund et al. (2000), Welch and King (2001), Dai et al. (2004), Srinivasan and Plattner (2006), Zia et al. (2007), Saeki et al. (2009), and Rafn et al. (2012) among many others.

4.2.2 Maize genetic data

In this section, we consider data on maize genetic properties. The data from maize genetic nested association mapping population discussed by McMullen et al. (2009) is publicly available and downloaded from http://www.panzea.org/lit/data_sets.html. The data contains 4699 samples or recombinant inbred lines combined across 25 families, representing 1106 SNP loci or genetic markers. For simplicity, to infer the genetic markers graph we treat the 4699 samples as replicates. The phenotypic variation measurements considered all reported by ordinal scale. Our objective is to identify trans-acting interactions of genetic markers across chromosomes in maize genome. The maize genome has 10 chromosomes. Trans-acting interactions also referred to as long-range chromosomal interactions or inter-chromosomal interactions has been studied, for example, in Miele and Dekker (2008) Lum and Merritt (2011).

We applied copula skeptic glasso to the maize genetics data. Using minimum BIC criterion we obtained the value of the tuning parameter close to 0.05 taken in the range 0.03 (dense) to 0.20 (chromosome-wise separated) The resulting network is displayed in Figure 4. As expected we see from Figure 4 that genetic markers within a chromo-

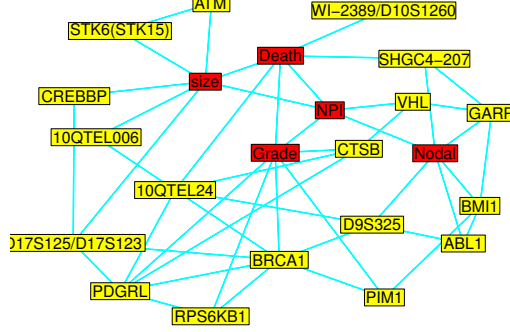


Figure 3: *Conditional dependence subgraph of clinical variables and selected genes from the breast cancer data. Red color or shaded rectangles represent clinical variables and yellow color or light shaded rectangles represent genomic variables.*

some are highly associated. On the other hand we see that some genetic markers form links across chromosomes. These potential links, for example, are between PZA01601.1 (chromosome 8) and PZA02480.1 (chromosome 5); PZA00473.5 (chromosome 6) and PZA03624.1 (chromosome 7); PZA02191.1 (chromosome 1) and PZA03321.4 (chromosome 2). These could refer trans-acting genetic markers which generate several interesting hypothesis for further experimental verifications. In support of our finding, McMullen et al. (2009) has also reported that among millions of pair-wise tests based on linkage disequilibrium (LD) marginally significant LD was observed between chromosome 6 and 7, though they concluded that it is a trivially small effect. We note that this could be possible in particular when a very large number of genetic markers are compared pair-wise, detecting even a single significant pair-wise association is often hard because of the large multiple testing adjustment factor involved (see also Bühlmann et al. (2014)). The graphical modeling approach presented in this paper can be an efficient tool towards the study of interactions of genetic markers within (cis-acting) and across (trans-acting) chromosomes.

5 Concluding Remarks

Large high-dimensional datasets have become a common feature of many modern measurement techniques. In this article, we have presented a sparse copula Gaussian graphical model to infer networks from large high-dimensional data sets of arbitrary type. We

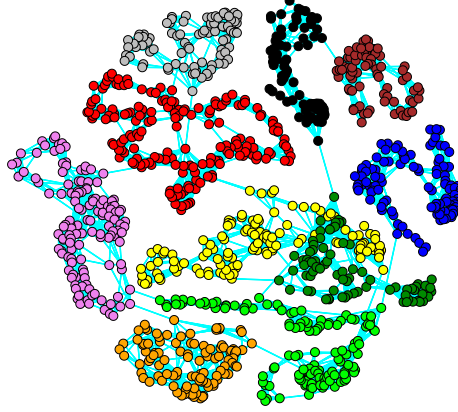


Figure 4: *Conditional dependence of genetic markers for the maize nested association mapping population. Genetic markers in each chromosome are represented by colours: red(Chr1), yellow(Chr2), green(Chr3), blue(Chr4), violet(Chr5), dark-green(Chr6), black(Chr7), orange(Chr8), grey(Chr9), and brown(Chr10)*

proposed two approaches for the analysis of high dimensional mixed variables: l_1 penalized extended rank likelihood Gaussian copula based EM algorithm and copula skeptic glasso with pairwise parametric copula selection, not necessarily from the same family.

The performance of the proposed approaches in comparison to existing methods are evaluated using simulation studies. The simulation results suggest that the proposed copula EM glasso and copula skeptic glasso perform well to identifying the true graph structure for high dimensional mixed variables setting. Taking into account computational efficiency we suggest to use the copula skeptic glasso for inferring networks for very high dimensional (thousands of variables) and copula EM glasso for moderately high dimensional mixed variables setting. Moreover, the EM copula glasso approach has the advantage that it can be directly implemented for missing data without any additional computational issue.

We have illustrated the application of the proposed graphical modeling approaches on gene amplifications and deletions microarray data from breast cancer experiment and genetic markers from the maize nested association mapping population. We obtained a sparse representation of the conditional dependencies between the clinical and genetic variables, which generated several interesting hypotheses on the importance of these variables for the treatment of breast cancer. In particular, we identified many genes that are amplified or deleted in breast cancer and may functionally contribute to ag-

gressiveness of breast cancer which is associated with worst outcome for the survival of breast cancer patients. The identification of such types of genes might lead to more accurate diagnostics and treatment at individual patient level. Similarly, a sparse representation of the interaction between genetic markers in maize genome, in particular across chromosomes will potentially be helpful for better understanding the molecular basis of phenotypic variation and to improve agricultural efficiency and sustainability. In general, the simulation and data analysis results suggest that the proposed copula based graphical modelings are promising approaches to infer networks for high dimensional nonnormal and mixed variables.

Bibliography

- C. Ambroise, J. Chiquet, and C. Matias. Inferring sparse gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3:205–238, 2009.
- M. Bärlund, F.z Forozan, Juha K., L. Bubendorf, Y. Chen, M. L Bittner, J. Torhorst, P. Haas, C. Bucher, G. Sauter, et al. Detecting activation of ribosomal protein s6 kinase by complementary dna and tissue microarray analysis. *Journal of the National Cancer Institute*, 92(15):1252–1259, 2000.
- Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view towards applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278, 2014.
- Q. Dai, Q. Y. Cai, X. O. Shu, A. Ewart-Toland, W. Q. Wen, A. Balmain, Y. T. Gao, and W. Zheng. Synergistic effects of stk15 gene polymorphisms and endogenous estrogen exposure in the risk of breast cancer. *Cancer Epidemiology Biomarkers & Prevention*, 13(12):2065–2070, 2004.
- A.P. Dempster. Covariance selection. *Biometrics*, pages 157–175, 1972.
- A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.
- A. Dobra and A. Lenkoski. Copula gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, 5(2A):969–993, 2011.
- J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432, 2008.
- C. Genest, K. Ghoudi, and L.P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552, 1995.

- P.J. Green. On use of the em for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 443–452, 1990.
- P.D. Hoff. Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, pages 265–283, 2007.
- J.G. Ibrahim, H. Zhu, and N. Tang. Model selection criteria for missing-data problems using the em algorithm. *Journal of the American Statistical Association*, 103(484): 1648–1658, 2008.
- L. B. Jensen, J.M.S. Bartlett, C.J. Witton, T. Kirkegaard, S. Brown, S. Müller, F. Campbell, T.G. Cooke, and K.V. Nielsen. Frequent amplifications and deletions of $g - \{1\}$ /s-phase transition genes, *ccnd1* and *myc* in early breast cancers: A potential role in $g - \{1\}$ /s escape. *Cancer Biomarkers*, 5(1):41–49, 2009.
- R. Kumar, A. Sharma, and R.K. Tiwari. Application of microarray in breast cancer: An overview. *Journal of Pharmacy & Bioallied Sciences*, 4(1):21, 2012.
- H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.
- Thomas E Lum and Thomas JS Merritt. Nonclassical regulation of transcription: Interchromosomal interactions at the malic enzyme locus of *drosophila melanogaster*. *Genetics*, 189(3):837–849, 2011.
- M. D McMullen, S. Kresovich, H. Villeda, H. Bradbury, P. and Li, Q. Sun, S. Flint-Garcia, J. Thornsberry, C. Acharya, C. Bottoms, et al. Genetic properties of the maize nested association mapping population. *Science*, 325(5941):737–740, 2009.
- Adriana Miele and Job Dekker. Long-range chromosomal interactions and gene regulation. *Molecular bioSystems*, 4(11):1046–1057, 2008.
- R. B Nelsen. *An introduction to copulas*. Springer, 2006.
- B. Rafn, C. F. Nielsen, S.H. Andersen, P. Szyniarowski, E. Corcelle-Termeau, E. Valo, N. Fehrenbacher, C.J. Olsen, M. Daugaard, C. Egebjerg, et al. Erbb2-driven breast cancer cell invasion depends on a complex signaling network activating myeloid zinc finger-1-dependent cathepsin b expression. *Molecular Cell*, 45(6):764–776, 2012.
- M. Ringnér, E. Fredlund, J. Häkkinen, Å. Borg, and J. Staaf. Gobo: Gene expression-based outcome for breast cancer online. *PloS one*, 6(3):e17911, 2011.
- M. Saeki, D. Kobayashi, N. Tsuji, K. Kuribayashi, and N. Watanabe. Diagnostic importance of overexpression of *bmi-1* mrna in early breast cancers. *International journal of oncology*, 35(3):511, 2009.
- A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8(1):11, 1959.

- D. Srinivasan and R. Plattner. Activation of abl tyrosine kinases promotes invasion of aggressive breast cancer cells. *Cancer research*, 66(11):5648–5655, 2006.
- C. Van Putten and J.H. Van Schuppen. Invariance properties of the conditional independence relation. *The Annals of Probability*, 13(3):934–945, 1985.
- P. L. Welch and M. C. King. Brca1 and brca2 and the genetics of breast and ovarian cancer. *Human Molecular Genetics*, 10(7):705–713, 2001.
- S. Wilhelm and B.G. Manjunath. tmvtnorm: A package for the truncated multivariate normal distribution. *sigma*, 2:2, 2010.
- M. K. Zia, K. A. Rmali, G. Watkins, R. E. Mansel, and W. G. Jiang. The expression of the von hippel-lindau gene product and its impact on invasiveness of human breast cancer cells. *International journal of molecular medicine*, 20(4):605, 2007.